Standard operating procedure (version 2.0) for RNASeq data processing QC in Yerkes Functional Genomics core

1 Short read mapping (RNA-seq reads from Illumina)

1.1 Checkout transcriptome annotation

Check out host, parasite, and spike-in control references, or previously used composite references from iRODS, after verifing previously checked out versions have not been updated)

1.2 Build composite reference genome STAR index

Convert annotations specified in ".gff" format to ".gtf" format with "gffread" for use in composite index building. Gene features are identified by the "gene_id" field in the ".gtf" annotation.

Example command:

```
gffread –T –o XXXRfGINXXCyXXXX_AnnotatedFeatures_XXXXXXX.gtf
XXXRfGINXXCyXXXX_AnnotatedFeatures_XXXXXXX.gff
```

Combine all the ".gtf" annotations to be used by the composite reference into one ".gtf" file with "cat".

Example command:

```
cat RfG7-8INMmXXXXXX_AnnotatedFeatures_XXXXXXX.gtf
XXXRfGINXXCyXXXX_AnnotatedFeatures_XXXXXXX.gtf ERCC92.gtf > Mm7-
8_Cy2-0_ERCC92.gtf
```

Build STAR index. STAR can take the sequence ".fasta" files from each reference in the composite as a separate argument and the composite annotation ".gtf" as one argument. The output is a STAR index file set used for mapping to the composite reference.

Example command:

```
STAR --runMode genomeGenerate --genomeDir Mm7-8_Cy2-0_ERCC92_S2.5.2b --
genomeFastaFiles RfG7-8INMmXXXXXX_Genome_XXXXXXX.fa
Pcyn_version_2_0/XXXRfGINXXCyXXXX_Genome_XXXXXXX.fa ERCC92.fa --
runThreadN 16 --sjdbGTFfile Mm7-8_Cy2-0_ERCC92.gtf --sjdbOverhang 100
```

1.3 Map the reads with STAR

All the “.fastq” files from each library (one library per sample) are mapped to a combined reference genome index using STAR version 2.5.2b. Options are set to include all multi-mapped alignments and unmapped reads so that the alignment status of every sequenced read is represented in the output alignment “.bam” file. STAR Gene counting is used to produce a read count abundance estimate for each “gene_id” in the composite annotation.

example command:

```
STAR --genomeDir /yerkes-cifs/runs/STAR/MaHPIC/Mm7-8_Cy2-0_ERCC92_S5.2.5b
--genomeLoad LoadAndKeep --readFilesIn 2063483_S5_L001_R1_001.fastq.gz,
2063483_S5_L002_R1_001.fastq.gz, 2063483_S5_L003_R1_001.fastq.gz
2063483_S5_L001_R2_001.fastq.gz, 2063483_S5_L002_R2_001.fastq.gz,
2063483_S5_L003_R2_001.fastq.gz --readFilesCommand zcat --runThreadN 16 --
outStdBAM_Unsorted --outSAMtype BAM Unsorted --outSAMmode Full --
outSAMattributesAll --outSAMstrandField intronMotif --outSAMunmapped Within --
outFileNamePrefix 2063483_ --outFilterMultimapNmax 999 --outSAMprimaryFlag
AllBestScore --quantMode GeneCounts > 2063483_MmCy_S252b_unsorted.bam
```

The output for this step includes the unsorted “.bam” file, one line for each mapped read, and the “_ReadsPerGene.out.tab” file which contains the read counts mapping to each “gene_id” in the composite annotation.

Sort and index alignment files.

Example command:

```
samtools sort -m 2G -@ 16 2063483_MmCy_S252b_unsorted.bam
2063483_MmCy_S252b_sorted
```

```
samtools index 2063483_MmCy_S252b_sorted.bam
```

1.3 Mapping Evaluation

Evaluate and summarize mapping quality including numbers of reads that mapped or didn’t map, ratio of reads that map to host vs parasite, …

Mapping stats

Reads Per Gene stats

3 5’-to-3’ coverage uniformity

For various reasons, RNA-seq read depth coverage can decrease when moving from the 3’ end to the 5’ end of transcripts. Therefore, another important quality control step is to make sure that coverage across the length of transcripts does not drop off precipitously.

Use Picard Tools: CollectRnaSeqMetrics, evaluate host coverage.

example command:

```
java -Xmx4g -jar ~/Downloads/Picard/picard-tools-1.74/CollectRnaSeqMetrics.jar
REF_FLAT=RfG7-8INMmXXXXXX_AnnotatedFeatures_XXXXXXX.refFlat
INPUT=2063483_MmCy_S252b_sorted.bam
OUTPUT=2063483_Mm_RnaSeqMetrics.txt STRAND_SPECIFICITY=NONE
CHART_OUTPUT=2063483_Mm_RnaSeqMetrics.pdf
```

2 Deliverables

BAM: Alignment files to composite reference

Read Count: Raw read count tables (samples/genes)

Metadata: Sample ID, metadata from each processing step, Counts and host/parasite ratio tables.

Coverage QC: RnaSeqMetrics text report and pdf plots.